Nonparametric predictive inference for comparison of two diagnostic tests

Manal Alabdulhadi[a], Tahani Coolen-Maturi[b,*], Frank P.A. Coolen[c]

[a]*Department of Mathematics, Qassim University, Qassim, Saudi Arabia.*
[b]*Department of Mathematical Sciences, Durham University, Durham, DH1 3LE, UK*
[c]*Department of Mathematical Sciences, Durham University, Durham, DH1 3LE, UK*

Abstract

An important aim in diagnostic medical research is comparison of the accuracy of two diagnostic tests. In this paper, comparison of two diagnostic tests is presented using nonparametric predictive inference (NPI) for future order statistics. The tests are assumed to be applied on the same individuals from two groups, e.g. healthy and diseased individuals, or from three groups with a known ordering, e.g. adding a group of severely diseased individuals to the two group scenario. Our comparison is explicitly in terms of lower and upper probabilities for proportions of correctly diagnosed future individuals from each group, for a given total number of such individuals. We include in our comparison the possibility that it is more important to get a correct diagnosis for individuals from one group than from another group.

Keywords: Comparing two diagnostic tests; lower and upper probabilities; nonparametric predictive inference.

1. Introduction

The performances of two diagnostic tests are traditionally compared in terms of their respective specificities and sensitivities, or by using a summary of these such as the area under the receiver operating characteristic (ROC) curve, this area is often referred to as AUC, see e.g. Pepe (2003, p77) and Zhou et al (2002, p27). In this paper, we present nonparametric predictive inference (NPI) as an alternative method for the comparison of two diagnostic tests.

NPI is a frequentist statistical method that is explicitly aimed at using few modelling assumptions, enabled through the explicit focus on events involving future observations and the use of lower and upper probabilities to quantify uncertainty (Augustin et al, 2014; Coolen, 2011). NPI has been introduced for many application areas where the predictive nature of

*Corresponding author
Email addresses: manalhamd@hotmail.com (Manal Alabdulhadi), tahani.maturi@durham.ac.uk (Tahani Coolen-Maturi), frank.coolen@durham.ac.uk (Frank P.A. Coolen)

this method plays an important role, including reliability, survival analysis, operations research and finance (see www.npi-statistics.com for more information). Restricting attention to one future observation, NPI has been developed for diagnostic test accuracy considering different types of data. For example, Coolen-Maturi et al. (2012b) introduced NPI for diagnostic test accuracy with binary data, while Elkhafifi and Coolen (2012) presented NPI for diagnostic tests with ordinal data. Coolen-Maturi et al. (2012a, 2014) proposed NPI for two and three group ROC analysis with continuous data. The results in Elkhafifi and Coolen (2012) have been generalised by Coolen-Maturi (2017a) for three group ROC analysis with ordinal data. Coolen-Maturi (2017b) presented NPI for scenarios where two or more diagnostic tests are combined in order to improve the overall diagnostic accuracy. Recently, NPI has been applied for inference on reproducibility of hypothesis tests (Coolen and Alqifari, 2018; Coolen and Bin Himd, 2014), presenting an interesting frequentist inference solution to a long-standing problem in medical and other applications.

The main difference between the NPI approach and established inference methods in the literature is that inferences in NPI are explicitly in terms of a given number of future individuals. In this paper, we present NPI for comparing two diagnostic tests, assuming that the two tests are applied to the same individuals from two or three groups. For two groups, one can typically think about healthy and diseased individuals, while for three groups there may e.g. be an added category of severaly diseased individuals. While such an association to medical applications is used in this paper, the methods presented are widely applicable in many fields. The predictive nature of the NPI approach can be attractive for diagnostic tests as one tends to assess the quality of the diagnostic tests for a given number of future individuals.

This paper is organized as follows. In Section 2, we give an overview of NPI for future order statistics. Comparison of two diagnostic tests using NPI for future order statistics, is presented in Section 3 for two groups of individuals, and in Section 4 for three groups. Section 5 contains some concluding remarks.

2. NPI for future order statistics

NPI is a frequentist statistical framework based on Hill's assumption $A_{(n)}$ (Hill, 1968), which yields direct probabilities for one or more future observations, based on n observations for related random quantities. $A_{(n)}$ does not assume anything else and it can be considered as a post-data assumption related to exchangeability. Inferences based on $A_{(n)}$ are nonparametric and predictive, and can be considered appropriate if there is hardly any information or knowledge about the random quantities of interest, other than the n observations (Hill, 1988). $A_{(n)}$ does not provide precise probabilities for many events of interest, however it provides bounds for all probabilities, these are lower and upper probabilities in the theory of interval probability (Augustin and Coolen, 2004; Weichselberger, 2000).

The assumption $A_{(n)}$ partially specifies a predictive probability distribution for one future observation. Suppose that $X_1, \ldots, X_n, X_{n+1}$ are continuous, real-valued and exchangeable random quantities. Suppose that the ordered observations of $X_1, \ldots, X_n$ are denoted by $x_1 < x_2 < ... < x_n$, and define $x_0 = -\infty$ and $x_{n+1} = \infty$ for ease of notation (or define

$x_0 = 0$ when dealing with non-negative random quantities). These n observations partition the real-line into $n+1$ intervals $I_j = (x_{j-1}, x_j)$, for $j = 1, 2, \ldots, n+1$. The assumption $A_{(n)}$ implies that the future observation X_{n+1} is equally likely to fall in any of these intervals with probability $\frac{1}{n+1}$ (Coolen, 2011). In NPI uncertainty is quantified by lower and upper probabilities for events of interest. Augustin and Coolen (2004) introduced predictive lower and upper probabilities based on $A_{(n)}$ as follows: Lower probability $\underline{P}(X_{n+1} \in B)$ and upper probability $\overline{P}(X_{n+1} \in B)$ for the event $X_{n+1} \in B$, based on the intervals $I_j = (x_{j-1}, x_j)$ $(j = 1, 2, \ldots, n+1)$ created by n real-valued non-tied observations, and the assumption $A_{(n)}$, are

$$\underline{P}(X_{n+1} \in B) = \frac{1}{n+1} \sum_j \mathbf{1}\{I_j \subseteq B\},$$
$$\overline{P}(X_{n+1} \in B) = \frac{1}{n+1} \sum_j \mathbf{1}\{I_j \cap B \neq \emptyset\}.$$

The NPI lower probability $\underline{P}(X_{n+1} \in B)$ is derived by taking only probability mass into account that is necessarily within B, which is only the case for the probability mass $\frac{1}{n+1}$ per interval I_j if this interval is completely contained within B. The upper probability $\overline{P}(X_{n+1} \in B)$ is achieved by taking all probability mass into account that could possibly be within B, which is the case for the probability mass $\frac{1}{n+1}$, per interval I_j, if the intersection of I_j and B is non-empty. Note that there are no further assumptions on the distribution of the probability mass $\frac{1}{n+1}$ in each interval I_j, so these lower and upper probabilities are the maximum lower and minimum upper bound, respectively, that can be derived for the event of interest. If one would make any further assumptions for the probability mass in each interval I_j, one would always end up with a (possibly imprecise) probability for the event $X_{n+1} \in B$ in between the NPI lower and upper probabilities.

We are interested in $m \geq 1$ future observations, X_{n+i} for $i = 1, \ldots, m$. We link the data and future observations via Hill's assumption $A_{(n)}$ (Hill, 1968), or more precisely, via $A_{(n+m-1)}$ (which implies $A_{(n+k)}$ for all $k = 0, 1, \ldots, m-2$), which can be considered as a post-data version of a finite exchangeability assumption for $n+m$ random quantities. $A_{(n+m-1)}$ implies that all possible orderings of the n data observations and the m future observations are equally likely, where the n data observations are not distinguished among each other, and neither are the m future observations. Let $S_j = \#\{X_{n+i} \in I_j,\ i = 1, \ldots, m\}$, then assuming $A_{(n+m-1)}$ we have

$$P(\bigcap_{j=1}^{n+1} \{S_j = s_j\}) = \binom{n+m}{n}^{-1}, \tag{1}$$

where s_j are non-negative integers with $\sum_{j=1}^{n+1} s_j = m$. Let $X_{(r)}$, for $r = 1, \ldots, m$, be the r-th ordered future observation, so $X_{(r)} = X_{n+i}$ for one $i = 1, \ldots, m$ and $X_{(1)} < X_{(2)} < \ldots < X_{(m)}$. The following probability is derived by counting the relevant orderings, for

$j = 1, \ldots, n+1$, and $r = 1, \ldots, m$,

$$P(X_{(r)} \in I_j) = \binom{j+r-2}{j-1} \binom{n-j+1+m-r}{n-j+1} \binom{n+m}{n}^{-1}. \quad (2)$$

For this event NPI provides a precise probability, as each of the $\binom{n+m}{n}$ equally likely orderings of n past and m future observations has the r-th ordered future observation in precisely one interval I_j (Coolen et al., 2018). The event that the number of future observations in an interval (x_a, x_b), denoted by C_{x_a,x_b}, is greater than or equal to a particular value v, has the following precise probability (Alqifari, 2017),

$$P(C_{x_a,x_b} \geq v) = \sum_{i=v}^{m} \binom{n+m}{n}^{-1} \binom{b-a-1+i}{i} \binom{n-b+a+m-i}{m-i}. \quad (3)$$

We use these NPI results for future order statistics in the comparison of diagnostic tests in this paper. For other applications of NPI for future order statistics we refer the reader to the recent PhD thesis by Alqifari (2017) and the related paper by Coolen et al. (2018).

3. Comparison of two diagnostic tests for two groups

We compare the accuracies of two diagnostic tests by explicitly considering the application of these tests to multiple future individuals. We assume that both diagnostic tests are applied to the same people. Assume that we have real-valued data from two different diagnostic tests on individuals from two groups in each test, say a 'healthy group' X and a 'disease group' Y, and there are n_x observations from the healthy group and n_y observations from the disease group. In our notation we indicate to the two tests by using superscript t; $t = 1, 2$, so we assume that we have data $(x_i^1, x_i^2), i = 1, ...n_x$ and $(y_j^1, y_j^2), j = 1, ...n_y$, where superscript 1 indicates test results of diagnostic test one and 2 indicates test results of diagnostic test two. We assume throughout this paper that the outcomes of the two tests are independent random quantities, given the disease states of the individuals, and that each individual had undergone both tests.

This section presents the comparison between two diagnostic tests for m_x and m_y future individuals. A natural question is whether one test is better than the other for the m_x and m_y future individuals from groups X and Y, respectively, and we investigate the possible influence of the choice of m_x and m_y, which mostly we will assume to be equal ($m_x = m_y = m$). This paper present the NPI method for comparison of two diagnostic tests with real-valued outcomes. It can be applied for any diagnostic tests, no matter which criterion is used to select the test threshold. In line with our recent work on diagnostic tests from NPI perspective, in our examples we will use the 2-NPI-L method for selecting the optimal threshold for each diagnostic test (Alabdulhadi, 2018; Coolen-Maturi et al., 2020).

Assume that we have real-valued data from two diagnostic tests on individuals from two groups. There are n_x observations from the healthy group X and n_y observations from the disease group Y. The ordered data from test t for groups X and Y are denoted by $x_1^t < x_2^t < \ldots < x_{n_x}^t$ and $y_1^t < y_2^t < \ldots < y_{n_y}^t$, respectively. For ease of presentation, we define

$x_0^t = y_0^t = -\infty$ and $x_{n_x+1}^t = y_{n_y+1}^t = \infty$. These n_x observations for test t applied to group X partition the real-line into $n_x + 1$ intervals $I_i^{X^t} = (x_{i-1}^t, x_i^t)$, for $i = 1, 2, \ldots, n_x + 1$, and the n_y observations for test t applied to group Y partition the real-line into $n_y + 1$ intervals $I_j^{Y^t} = (y_{j-1}^t, y_j^t)$, for $j = 1, \ldots, n_y + 1$. In this section, we consider m_x future individuals from group X, with random results from diagnostic test t denoted by $X_{n_x+r}^t$, $r = 1, \ldots, m_x$, and m_y future individuals from group Y, with random results from diagnostic test t denoted by Y_{n_y+s}, $s = 1, \ldots, m_y$. We will particularly use the ordered future observations, we denote the m_x and m_y ordered future observations from groups X and Y, for test t, by $X_{(1)}^t < X_{(2)}^t < \ldots < X_{(m_x)}^t$ and $Y_{(1)}^t < Y_{(2)}^t < \ldots < Y_{(m_y)}^t$, respectively.

We assume that small values of the diagnostic test results are associated with absence of the disease and large values of the test results with presence of the disease. Of course, the method can directly be applied if this were the other way around, but we restrict attention to tests with such a simple pattern, hence which require only a single threshold for the diagnosis. However, in Section 4 we consider a scenario with one 'middle group' and two thresholds, this indicates how more complicated tests with two groups but which require multiple thresholds, e.g. if a finite interval of values indicate 'healthy' while being outside this interval on either side indicates 'disease', can also be studied in a similar way. So we assume that a threshold $c^t \in \mathbb{R}$ is used to classify individuals to either being healthy if their test result is less than or equal to c^t or having the disease if their test result is greater than c^t.

For a specific value of c^t, $C_{c^t}^{X^t}$ denotes the number of correctly classified future individuals from the healthy group X by test t, that is those with test results $X_{n_x+r}^t \leq c^t$ (for $r = 1, \ldots, m_x$), and $C_{c^t}^{Y^t}$ denotes the number of correctly classified future individuals from the disease group Y by test t, that is those with test results $Y_{n_y+s}^t > c^t$ (for $s = 1, \ldots, m_y$). Let α and β be any two values in $(0, 1]$ that are selected to reflect the desired quality of the diagnoses and the importance of correct diagnosis for one group compared to correct diagnosis for the other group. We consider the aim that the number of correctly classified individuals out of m_x future individuals from the healthy group X, is at least αm_x, and that the number of correctly classified individuals out of m_y future individuals from the disease group Y is at least βm_y.

Using the independence assumption between the two groups, the joint NPI lower and upper probabilities for the event of interest can be derived as the products of the corresponding NPI lower and upper probabilities for the individual events that involve $C_{c^t}^{X^t}$ and $C_{c^t}^{Y^t}$, so

$$\underline{P}(C_{c^t}^{X^t} \geq \alpha m_x, C_{c^t}^{Y^t} \geq \beta m_y) = \underline{P}(C_{c^t}^{X^t} \geq \alpha m_x) \times \underline{P}(C_{c^t}^{Y^t} \geq \beta m_y), \tag{4}$$

$$\overline{P}(C_{c^t}^{X^t} \geq \alpha m_x, C_{c^t}^{Y^t} \geq \beta m_y) = \overline{P}(C_{c^t}^{X^t} \geq \alpha m_x) \times \overline{P}(C_{c^t}^{Y^t} \geq \beta m_y). \tag{5}$$

We use the NPI results for future order statistics, reviewed in Section 2, to derive the NPI lower and upper probabilities in Equations (4) and (5). We first present the results for group X, followed by those for group Y, for which deriving the results follows similar steps.

Note that the event $C_{c^t}^{X^t} \geq \alpha m_x$ is equivalent to the event $X_{(\lceil \alpha m_x \rceil)}^t \leq c^t$, where $\lceil \alpha m_x \rceil$ denotes the smallest integer greater than αm_x, and similar for group Y.

We introduce the notation $I_i^{X^t} = (x_{i-1}^t, x_i^t)$ for $i = 1, \ldots, n_x+1$ and let $i_{c^t}^x \in \{1, 2, \ldots, n_x+1\}$ be such that $c^t \in I_{i_{c^t}}^{X^t} = (x_{i_{c^t}^x - 1}^t, x_{i_{c^t}^x}^t)$. The NPI lower and upper probabilities for the event $C_{c^t}^{X^t} \geq \alpha m_x$ are

$$\underline{P}(C_{c^t}^{X^t} \geq \alpha m_x) = \underline{P}(X_{(\lceil \alpha m_x \rceil)} \leq c^t) = \sum_{i=1}^{i_{c^t}^x - 1} P(X_{(\lceil \alpha m_x \rceil)} \in I_i^{X^t}), \tag{6}$$

$$\overline{P}(C_{c^t}^{X^t} \geq \alpha m_x) = \overline{P}(X_{(\lceil \alpha m_x \rceil)} \leq c^t) = \sum_{i=1}^{i_{c^t}^x} P(X_{(\lceil \alpha m_x \rceil)} \in I_i^{X^t}), \tag{7}$$

where the precise probabilities on the right hand sides of Equations (6) and (7) can be obtained from Equation (2).

The NPI lower and upper probabilities for the event $C_{c^t}^{Y^t} \geq \beta m_y$ are derived similarly. Introducing notation $I_j^{Y^t} = (y_{j-1}^t, y_j^t)$ for $j = 1, \ldots, n_y + 1$ and letting $j_{c^t}^y \in \{1, 2, \ldots, n_y + 1\}$ be such that $c^t \in I_{j_{c^t}^y}^{Y^t} = (y_{j_{c^t}^y - 1}^t, y_{j_{c^t}^y}^t)$, we have

$$\underline{P}(C_{c^t}^{Y^t} \geq \beta m_y) = \underline{P}(Y_{(m_y - \lceil \beta m_y \rceil + 1)} > c^t) = \sum_{j=j_{c^t}^y + 1}^{n_y+1} P(Y_{(m_y - \lceil \beta m_y \rceil + 1)} \in I_j^{Y^t}), \tag{8}$$

$$\overline{P}(C_{c^t}^{Y^t} \geq \beta m_y) = \overline{P}(Y_{(m_y - \lceil \beta m_y \rceil + 1)} > c^t) = \sum_{j=j_{c^t}^y}^{n_y+1} P(Y_{(m_y - \lceil \beta m_y \rceil + 1)} \in I_j^{Y^t}). \tag{9}$$

We use the NPI lower and upper probabilities from Equations (4) and (5) to compare two diagnostic tests. We consider it a strong indication that test 1 is better than test 2 if

$$\underline{P}(C_{c^1}^{X^1} \geq \alpha m_x, C_{c^1}^{Y^1} \geq \beta m_y) > \overline{P}(C_{c^2}^{X^2} \geq \alpha m_x, C_{c^2}^{Y^2} \geq \beta m_y). \tag{10}$$

We further consider it a weak indication for test 1 being better than test 2 if both

$$\underline{P}(C_{c^1}^{X^1} \geq \alpha m_x, C_{c^1}^{Y^1} \geq \beta m_y) > \underline{P}(C_{c^2}^{X^2} \geq \alpha m_x, C_{c^2}^{Y^2} \geq \beta m_y), \tag{11}$$

and

$$\overline{P}(C_{c^1}^{X^1} \geq \alpha m_x, C_{c^1}^{Y^1} \geq \beta m_y) > \overline{P}(C_{c^2}^{X^2} \geq \alpha m_x, C_{c^2}^{Y^2} \geq \beta m_y). \tag{12}$$

Next, the NPI method for comparison of two diagnostic tests with two groups, as introduced above, is illustrated in an example which is created to provide insight into the method, followed by an example showing the application of the method with data from the literature. For more exploration of the method through further examples we refer to the PhD thesis of the first-named author (Alabdulhadi, 2018).

Example 3.1. Consider an artificial data set from two different diagnostic tests applied to the same individuals from two groups, containing the test results from $n_x = n_y = 10$ individuals from each group. Note that our nonparametric method effectively works with ranks, hence for these first two illustrative examples we use integer data that could be interpreted as ranks. For test 1, the data for the healthy group, are $\{1, 2, 3, 4, 5, 7, 9, 10, 11, 12\}$ and for the disease group the data are $\{6, 8, 13, 14, 15, 16, 17, 18, 19, 20\}$. Note that in the general presentation of our method above, these data were denoted by x_i^1 and y_j^1, for $i, j = 1, 2, \ldots, 10$, respectively. For test 2, the data for the healthy group are $\{1, 2, 6, 7, 10, 11, 12, 13, 16, 18\}$ and for the disease group we have the observations $\{3, 4, 5, 8, 9, 14, 15, 17, 19, 20\}$. These data clearly suggest that test 1 differentiates groups X and Y better than test 2. To determine the thresholds used for the tests, we applied the 2-NPI-L method (Alabdulhadi, 2018; Coolen-Maturi et al., 2020), aimed at optimal diagnostic performance by actually choosing the thresholds in order to maximize the NPI lower probability of interest, so as presented in Equation (4). With $\alpha = \beta = 0.6$, this resulted in optimal thresholds $c^1 \in (12, 13)$ for test 1 and $c^2 \in (13, 14)$ for test 2, for all the considered values of m. Applying these thresholds to the empirical data would lead to all 10 healthy people and 8 of the 10 diseased people being correctly diagnosed for test 1, while test 2 would provide the correct diagnosis for 8 of the 10 healthy people and 5 of the 10 diseased people. It should be noted that, for the novel method to compare two diagnostic tests presented in this paper, the method used to determine the thresholds is irrelevant, hence we do not pay more attention to this here.

The NPI lower and upper probabilities given by Equations (4) and (5) for test 1 and test 2, where we consider the same number $m = m_x = m_y$ future observations for both the healthy and disease groups, with the results for $m = 1, 2, \ldots, 30$ presented in Figure 1. The values $\alpha = \beta = 0.6$ are used for the left plot and $\alpha = 0.9, \beta = 0.1$ for the right plot. The left plot, for $\alpha = \beta = 0.6$, shows that test 1 quite easily leads to a successful test according to the criterion that at least 60% of future individuals from each of the two groups should be correctly diagnosed, as the NPI lower probabilities for this event (the left end-point of the plotted intervals), for the different values of m, are quite large and the corresponding upper probabilities (the right end-points) are mostly close to 1. The results show quite some variability for small values of m, this is due to the discrete nature of the number of required future successes. In particular, for $m = 2$ it is harder to achieve the criterion of at least 60% correct diagnoses for both groups than for $m = 1$. This effect becomes smaller for larger values of m, hence there is less variation in the results for larger m. So, for $\alpha = \beta = 0.6$, our method provides a strong indication that test 1 performes better than test 2 for all considered values of m.

We have included the second case, with $\alpha = 0.9, \beta = 0.1$ to illustrate our method in a scenario where it is very important to get the correct diagnosis for the healthy group X but hardly relevant for the disease group Y. Of course, this is probably quite an unrealistic scenario, and in practice one could even neglect the results for group Y, but we include it to illustrate some further features of our method. The optimal threshold, according to the NPI lower method (Alabdulhadi, 2018; Coolen-Maturi et al., 2020), is now again $c^1 \in (12, 13)$

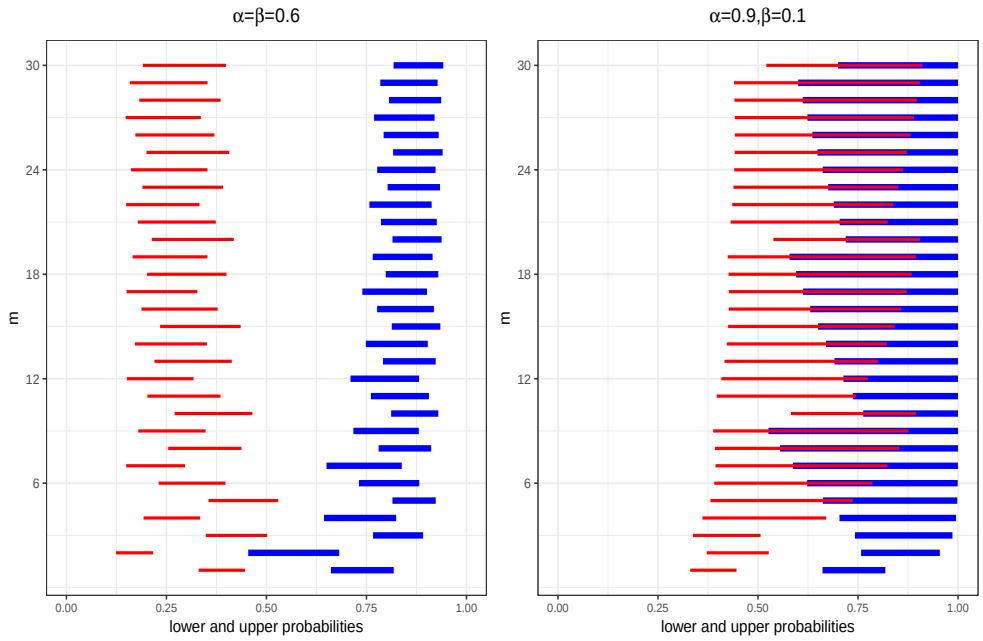

Figure 1: Comparison of Test 1 (blue, thick) and Test 2 (red, thin).

for test 1, for all considered values of m. This is logical as these are the smallest values of c^1 for which all empirical observations from group X are correctly diagnosed. But for test 2 it is quite different from the scenario discussed above, as now the optimal threshold is $c^2 \in (13, 14)$ for $m = 1, 2, 3$ and $c^2 \in (18, 19)$ for $m = 4, \ldots, 30$.

The right plot in Figure 1 shows that, for small values of m, there remains a strong indication that test 1 is better than test 2, with this criterion emphasizing the importance to get the diagnoses for group X correct. However, for $m \in \{5, 6, \ldots, 30\}$, there is only a weak indication that test 1 is better than test 2. This illustrates that the specific value of m can affect the conclusion of the comparison of diagnostic tests.

Example 3.2. In this example, we use the data set from a study to develop screening methods to detect carriers of a rare genetic disorder. The data were discussed by Cox et al (1982), and are available from Carnegie Mellon University Statlib Datasets Archive (http://lib.stat.cmu.edu/datasets/). Four different tests were used for each patient, denoted by T^{M1}, T^{M2}, T^{M3} and T^{M4}. For some patients, there are several samples of which the average is considered, and five missing values are excluded from the analysis. The remaining sample, which is used in this example, consists of 120 observations, 38 for carriers of the rare genetic disorder, which is the disease group Y in our terminology, and 82 for non-carriers, the healthy group X. In this example, we use this data set for pairwise comparisons of these four diagnostic tests, using the NPI method presented in this section.

To compare any two of these four tests, the 2-NPI lower and upper probabilities as given in Equations (4) and (5), for $m = 1, \ldots, 30$, are presented in Figures 2 and 3, for the

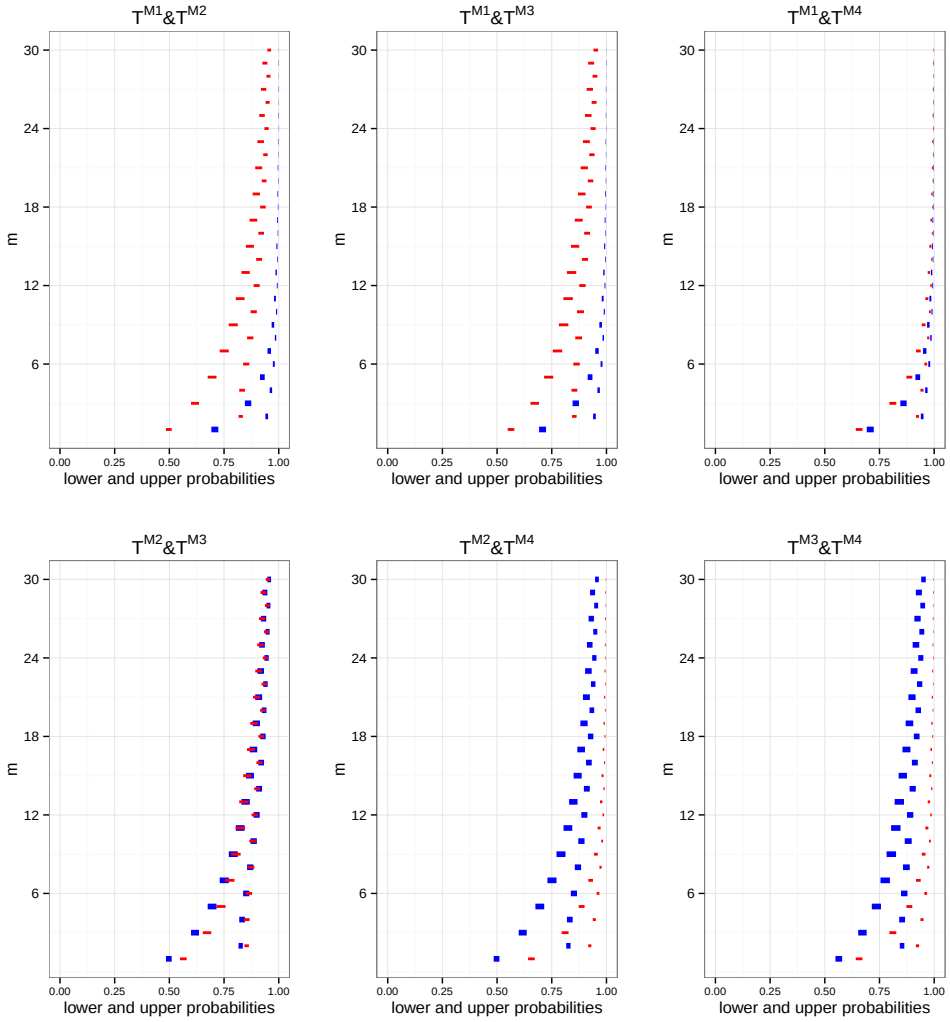

Figure 2: Pairwise comparisons of T^{M1}, T^{M2}, T^{M3} and T^{M4}, with $\alpha = \beta = 0.5$.

scenarios $\alpha = \beta = 0.5$ and $\alpha = 0.5, \beta = 0.7$, respectively. The heading of each plot states the two diagnostic tests for which these lower and upper probabilities are presented by intervals with these values as endpoints, the first named test is presented in blue (thick line) and the second named test in red (thin line).

The optimal threshold for each test has again been determined by the NPI lower method (Alabdulhadi, 2018; Coolen-Maturi et al., 2020), which corresponds to the NPI method for comparison of diagnostic tests as presented in this paper. The optimal thresholds lead to the following numbers of correctly classified individuals in the data. For $\alpha = \beta = 0.5$, test T^{M1} correctly classifies 70 out of 82 individuals from group X and 32 out of 38 from group Y, for all $m = 1, \ldots, 30$. For test T^{M2} the numbers are 56 out of 82 for group X and 28 out of 38 for group Y for $m = 1, 2$, while for all $m = 3, \ldots, 30$ the numbers are 58 out of 82 from group X and 27 out of 38 from group Y. Test T^{M3} classifies 74 out of 82 of the data from

group X correctly, together with 24 out of 38 from group Y when the optimal threshold for the case $m = 1$ is used. If $m = 2, \ldots, 11$ these numbers change to 70 out of 82 from group X and 25 out of 38 from group Y. If $m = 12, \ldots, 30$, test T^{M3} classifies 57 out of 82 data values from group X correctly, together with 27 out of 38 from group Y. Test T^{M4} leads to the same optimal threshold for all considered values of m, namely 67 out of 82 from group X and 31 out of 38 from group Y. So, the empirical results for these tests indicate that the numbers of correctly classified individuals from both groups for the largest for test T^{M1}, followed by the numbers for test T^{M4}. If we consider $m = 1, \ldots, 11$ then the numbers of correctly classified individuals from both groups for T^{M3} are greater than the corresponding numbers for T^{M2}. But for $m = 12, \ldots, 30$, the number of correctly classified individuals from group X is greater for T^{M2} than for T^{M3} while the numbers from group Y are equal for these two tests. We have included this discussion of the empirical performances of these tests as they are also reflected in the predictive performances as used in the NPI method to compare diagnostic tests presented in this paper.

For the case $\alpha = \beta = 0.5$, presented in Figure 2, our NPI method to compare two diagnostic tests leads to the following conclusions. The first two plots show that there is a strong indication that test T^{M1} is better than both tests T^{M2} and T^{M3} for all considered values of m. The third plot shows that for the larger values of m there is a weak indication that test T^{M1} is better than T^{M4}, but for smaller values of m there is a strong indication that T^{M1} is better than T^{M4}. Note that both these tests clearly do well on satisfying the predictive criterion of classifying at least half the future people from both the healthy and disease groups correctly. From the first plot in the second row we notice that there is either a strong or weak indication that T^{M3} is better than T^{M2} for smaller values of m, whereas for larger values of m are nested there is a weak indication that T^{M2} performs better than T^{M3}. This is of course in line with the above discussed empirical performances of these two methods, and it clearly illustrates that conclusions on comparative predictive performances of two diagnostic tests can depend on the number of future individuals considered. The final two plots show that there is a strong indication that T^{M4} is better than both T^{M2} and T^{M3} for all considered values of m. It should be noted that the imprecision, that is the difference between corresponding upper and lower probabilities, is smaller in this example than in Example 3.1. This is due to the fact that we have considerably more data in this example leading to reduced imprecision.

We also consider our method with the required proportions of correctly classified individuals per group set at $\alpha = 0.5$ and $\beta = 0.7$. With these values, the optimal thresholds vary a bit more for the tests than in the case discussed above. Applying the NPI lower method (Alabdulhadi, 2018; Coolen-Maturi et al., 2020) to determine the optimal diagnostic threshold, leads to the following numbers of correctly classified individuals from the data set. For test T^{M1}, 70 out of 82 from group X and 32 out of 38 from group Y are correctly classified for $m = 1, 2, 4, 5, 7, 9, 11$, while for $m = 6, 8, 12, 13, 16, 18, 22, 23, 24, 26, 28, 29, 30$ the numbers are 56 out of 82 from group X and 34 out of 38 from group Y, and for $m = 3, 10, 14, 15, 17, 19, 20, 21, 25, 27$ we have 60 out of 82 from group X and 33 out of 38 from group Y correctly classified. For test T^{M2}, for $m = 1, 5, 7$ there are 56 out of 82 correctly classified individuals for group X and 28 out of 38 for group Y, while for all other

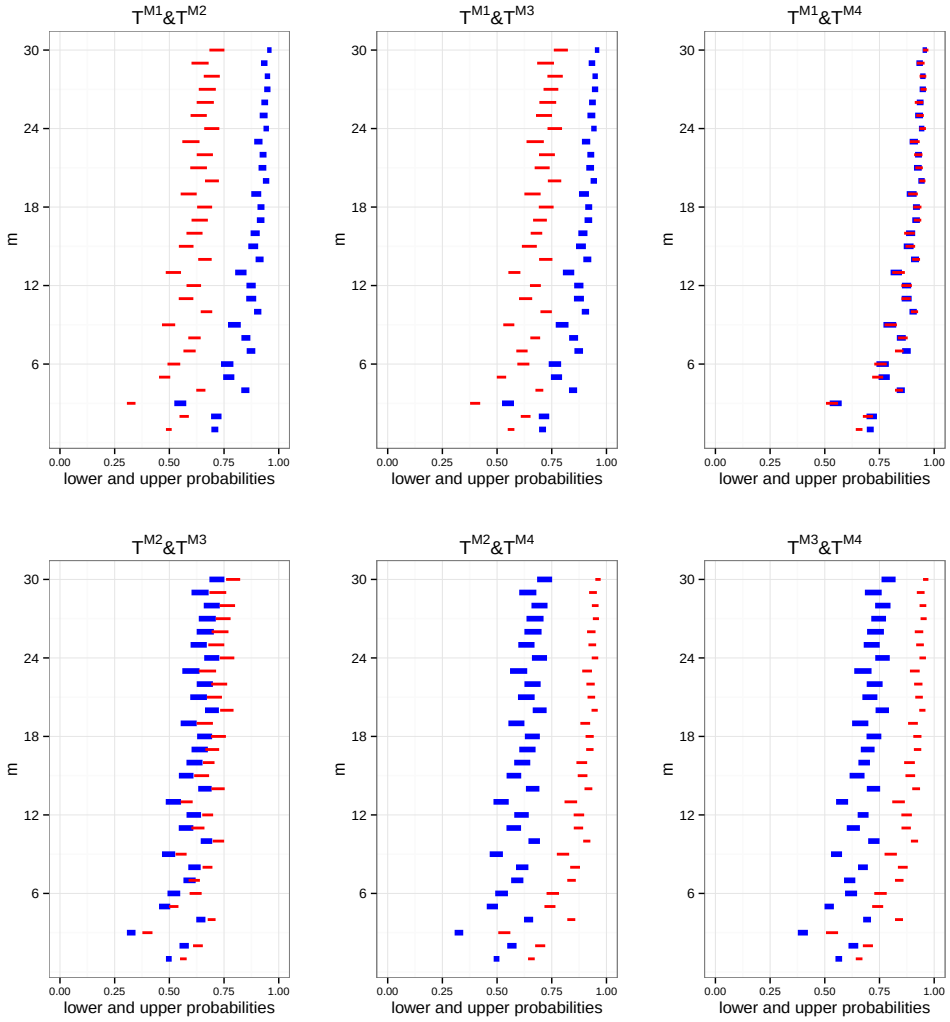

Figure 3: Pairwise comparisons of T^{M1}, T^{M2}, T^{M3} and T^{M4}, with $\alpha = 0.5, \beta = 0.7$.

values of m considered the numbers are 36 out of 82 and 35 out of 38. For test T^{M3}, the case $m = 1$ leads to 74 out of 82 correctly classified individuals in the data from group X and 24 out of 38 from group Y. For $m = 2, 3, 4, 5, 6, 8, 9, 12, 13, 16$, test T^{M3} leads to 42 out of 82 correctly classified data observations from group X and 35 out of 38 from group Y, while for $m = 7, 10, 11, 15, 17, \ldots, 30$ this test leads to 52 out of 82 correctly classified individuals from group X and 30 out of 38 from group Y. Finally, for test T^{M4} we have, for $m = 1$, 67 out of 82 correct classifications for the data from group X and 31 out of 38 from group Y, while for $m = 2, 6$ the numbers are 55 out of 82 from group X and 34 out of 38 from group Y. For all other considered valued of m, T^{M4} correctly classifies 61 out of 82 individuals in the data from group X and 33 out of 38 from group Y.

Our new NPI method for comparison of two diagnostic test is presented in the six plots

of pairwise comparisons in Figure 3 for this case with $\alpha = 0.5, \beta = 0.7$. The results are largely similar to those with the values $\alpha = \beta = 0.5$ as presented in Figure 2 and discussed above. Of course, the NPI lower and upper probabilities for the events of interest are now lower than before because the required number of correctly classified future individuals from group Y is larger that $\beta = 0.5$. The plots show somewhat more variation, which reflects the same increased variation in the results for the empirical data classifications discussed above. One of the small changes compared to the previous case can be seen from the third plot, as test T^{M1} is not always better anymore than T^{M4}, indeed there are now several values of m for which there is a weak indication that T^{M1} is better than T^{M4} and also several values of m for which the opposite is weakly indicated. Of course, considering the actual values of the NPI lower and upper probabilities in this comparison between T^{M1} and T^{M4} it is clear that they are very close.

4. Comparison of two diagnostic tests for three groups

In this section we extend the method of the previous section to comparison of two diagnostic tests with three ordered groups. Such a scenario occurs, for example, if the diseased individuals can be divided into two groups, with less and more severe levels of the disease, which may be relevant with regard to their treatments. We extend the notation introduced above by denoting this third group of most severely diseased individuals by Z, and we suppose to have n_z observed test results, for both tests considered, for members of this group. The ordered data from test t for this group are denoted by $z_1^t < z_2^t < \ldots < z_{n_z}^t$, and we define $z_0^t = -\infty$ and $z_{n_z+1}^t = \infty$. These n_z observations create the intervals $I_l^{Z^t} = (z_{l-1}^t, z_l^t)$, for $l = 1, 2, \ldots, n_z+1$. Let the diagnostic test results of test t applied to m_z future individuals be denoted by $Z_{n_z+k}^t$, $k = 1, \ldots, m_z$, and let the corresponding ordered future observations be denoted by $Z_{(1)}^t < Z_{(2)}^t < \ldots < Z_{(m_z)}^t$. Assume that the three groups are ordered in the sense that, for both tests considered, observations from group X tend to be smaller than those from group Y, which in turn tend to be smaller than those from group Z. For a diagnostic decision rule for test t, two thresholds $c_1^t < c_2^t$ are required to classify individuals into one of the three groups. If test value, for test t, is less than or equal to c_1^t the diagnosis is that the individual belongs to group X. A test value which is greater than c_1^t and less than or equal to c_2^t is an indication that the individual belongs to group Y, while a test value greater than c_2^t leads to the diagnosis that the individual belongs to group Z. The numbers of correctly classified individuals, using test t, out of the m_x, m_y and m_z future individuals for groups X, Y and Z, are denoted by $C_{c_1^t}^{X^t}$, $C_{(c_1^t, c_2^t)}^{Y^t}$ and $C_{c_2^t}^{Z^t}$, respectively.

We also extend the criterion for successful diagnostic performance of a test by introducing $\gamma \in (0, 1]$ to denote the minimum proportion of future individuals from group Z which the test should diagnose correctly, in addition to the proportions α and β for groups X and Y. Assuming that the three groups are fully independent, in the sense that any information about individuals in one group does not provide any information about individuals in the other groups, the joint NPI lower and upper probabilities for the event of interest are

$$\underline{P}(C_{c_1^t}^{X^t} \geq \alpha m_x, C_{(c_1^t,c_2^t)}^{Y^t} \geq \beta m_y, C_{c_2^t}^{Z^t} \geq \gamma m_z) =$$
$$\underline{P}(C_{c_1^t}^{X^t} \geq \alpha m_x) \times \underline{P}(C_{(c_1^t,c_2^t)}^{Y^t} \geq \beta m_y) \times \underline{P}(C_{c_2^t}^{Z^t} \geq \gamma m_z), \tag{13}$$
$$\overline{P}(C_{c_1^t}^{X^t} \geq \alpha m_x, C_{(c_1^t,c_2^t)}^{Y^t} \geq \beta m_y, C_{c_2^t}^{Z^t} \geq \gamma m_z) =$$
$$\overline{P}(C_{c_1^t}^{X^t} \geq \alpha m_x) \times \overline{P}(C_{(c_1^t,c_2^t)}^{Y^t} \geq \beta m_y) \times \overline{P}(C_{c_2^t}^{Z^t} \geq \gamma m_z). \tag{14}$$

For $I_i^{X^t} = (x_{i-1}^t, x_i^t)$ with $i = 1, \ldots, n_x + 1$ and $c_1^t \in I_{i_{c_1^t}}^{X^t} = (x_{i_{c_1^t}-1}^t, x_{i_{c_1^t}}^t)$, $i_{c_1^t} \in \{1, \ldots, n_x + 1\}$, the NPI lower and upper probabilities for the event $C_{c_1^t}^{X^t} \geq \alpha m_x$ are

$$\underline{P}(C_{c_1^t}^{X^t} \geq \alpha m_x) = \underline{P}(X_{\lceil \alpha m_x \rceil} \leq c_1^t) = \sum_{i=1}^{i_{c_1^t}-1} P(X_{\lceil \alpha m_x \rceil} \in I_i^{X^t}), \tag{15}$$

$$\overline{P}(C_{c_1^t}^{X^t} \geq \alpha m_x) = \overline{P}(X_{\lceil \alpha m_x \rceil} \leq c_1^t) = \sum_{i=1}^{i_{c_1^t}} P(X_{\lceil \alpha m_x \rceil} \in I_i^{X^t}). \tag{16}$$

For $I_j^{Y^t} = (y_{j-1}^t, y_j^t)$ with $j = 1, \ldots, n_y + 1$ and $c_1^t \in I_{j_{c_1^t}}^Y = (y_{j_{c_1^t}-1}, y_{j_{c_1^t}})$ and $c_2^t \in I_{j_{c_2^t}}^Y = (y_{j_{c_2^t}-1}, y_{j_{c_2^t}})$, with $j_{c_1^t} \in \{1, \ldots, n_y + 1\}$ and $j_{c_2^t} \in \{1, \ldots, n_y + 1\}$, with $c_2^t \geq c_1^t$, which implies that $j_{c_2^t} \geq j_{c_1^t}$, we get NPI lower and upper probabilities

$$\underline{P}(C_{(c_1^t,c_2^t)}^{Y^t} \geq \beta m_y) = P(C_{(y_{j_{c_1^t}}, y_{j_{c_2^t}-1})}^{Y^t} \geq \beta m_y), \tag{17}$$

$$\overline{P}(C_{(c_1^t,c_2^t)}^{Y^t} \geq \beta m_y) = P(C_{(y_{j_{c_1^t}-1}, y_{j_{c_2^t}})}^{Y^t} \geq \beta m_y). \tag{18}$$

The probabilities on the right-hand sides of these equations can be computed using Equation (3).

For $I_l^{Z^t} = (z_{l-1}^t, z_l^t)$ with $l = 1, \ldots, n_z + 1$ and $c_2^t \in I_{l_{c_2^t}}^{Z^t} = (z_{l_{c_2^t}-1}^t, z_{l_{c_2^t}}^t)$, $l_{c_2^t} = 1, \ldots, n_z + 1$, the NPI lower and upper probabilities are

$$\underline{P}(C_{c_2^t}^{Z^t} \geq \gamma m_z) = \underline{P}(Z_{(m_z - \lceil \gamma m_z \rceil + 1)} > c_2^t) = \sum_{l=l_{c_2^t}+1}^{n_z+1} P(Z_{(m_z - \lceil \gamma m_z \rceil + 1)} \in I_l^{Z^t}), \tag{19}$$

$$\overline{P}(C_{c_2^t}^{Z^t} \geq \gamma m_z) = \overline{P}(Z_{(m_z - \lceil \gamma m_z \rceil + 1)} > c_2^t) = \sum_{l=l_{c_2^t}}^{n_z+1} P(Z_{(m_z - \lceil \gamma m_z \rceil + 1)} \in I_l^{Z^t}). \tag{20}$$

Similar to the NPI comparison of two diagnostic tests with two groups presented above, we compare the two diagnostic tests with three groups as follows. We consider it a strong indication that test 1 is better than test 2 if

$$\underline{P}(C_{c_1^1}^{X^1} \geq \alpha m_x, C_{(c_1^1,c_2^1)}^{Y^1} \geq \beta m_y, C_{c_2^1}^{Z^1} \geq \gamma m_z) >$$
$$\overline{P}(C_{c_1^2}^{X^2} \geq \alpha m_x, C_{(c_1^2,c_2^2)}^{Y^2} \geq \beta m_y, C_{c_2^2}^{Z^2} \geq \gamma m_z). \tag{21}$$

We consider it a weak indication that test 1 is better than test 2 if both

$$\underline{P}(C_{c_1^1}^{X^1} \geq \alpha m_x, C_{(c_1^1,c_2^1)}^{Y^1} \geq \beta m_y, C_{c_2^1}^{Z^1} \geq \gamma m_z) > \\ \underline{P}(C_{c_1^2}^{X^2} \geq \alpha m_x, C_{(c_1^2,c_2^2)}^{Y^2} \geq \beta m_y, C_{c_2^2}^{Z^2} \geq \gamma m_z), \tag{22}$$

and

$$\overline{P}(C_{c_1^1}^{X^1} \geq \alpha m_x, C_{(c_1^1,c_2^1)}^{Y^1} \geq \beta m_y, C_{c_2^1}^{Z^1} \geq \gamma m_z) > \\ \overline{P}(C_{c_1^2}^{X^2} \geq \alpha m_x, C_{(c_1^2,c_2^2)}^{Y^2} \geq \beta m_y, C_{c_2^2}^{Z^2} \geq \gamma m_z). \tag{23}$$

As before, our NPI method to compare two diagnostic tests can be applied for any thresholds used in the tests. For illustration of our method in the following example, which uses data from the literature, we have used the 3-NPI-L method that we recently introduced (Alabdulhadi, 2018; Coolen-Maturi et al., 2020). This method determines the thresholds c_1^t and c_2^t which maximise the joint NPI lower probability given in Equation (13).

Example 4.1. The interleukin-6 (IL-6) and serum soluble triggering receptor expressed (sTREM-1) are common diagnostic tests for detection of late onset sepsis (LOS) in neonates (Sarafidis et al, 2010). Both these diagnostic tests were applied to 52 neonates assessed as suspicious for LOS. They were classified into three groups, 21 non-infected neonates (no laboratory evidence of sepsis and negative blood cultures), 9 possible sepsis (laboratory evidence of sepsis however negative blood cultures) and 22 confirmed sepsis (positive blood cultures for fungi and microbes). We refer to these groups as X, Y and Z, respectively, they are logically ordered in sense of severeness of the disease to the individual.

The NPI lower and upper probabilities as given in Equations (13) and (14) for tests IL-6 and sTREM-1 are presented in Figure 4, for $m = 1, \ldots, 30$, we have used $m_x = m_y = m_z = m$ throughout this example. We have considered two different criteria expressed by the proportions α, β and γ. The 3-NPI-L method was applied to determine the two optimal thresholds for each test (Alabdulhadi, 2018; Coolen-Maturi et al., 2020). First we consider the numbers of correctly classified individuals in the data set, using these thresholds.

For $\alpha = \beta = \gamma = 0.5$, the numbers of correctly classified individuals from groups X, Y and Z for test IL-6 are 18 out of 21, 6 out of 9 and 11 out of 22, respectively, for all $m = 1, \ldots, 30$. For test sTREM-1 the corresponding numbers are 16 out of 21, 5 out of 9 and 6 out of 22, also for all considered values of m. So, with $\alpha = \beta = \gamma = 0.5$, test IL-6, applied to the available data, meets the criteria set on the proportions from each group that are correctly diagnosed, for each of the three groups X, Y and Z. However, test sTREM-1 does not correctly classify at least 50% of the data observations from group Z. This is reflected in the predictive results from our new NPI method, presented in the first plot in Figure 4, which provides a strong indication that test IL-6 (blue, thick line) is better than test sTREM-1 (red, thin line) for all $m = 1, \ldots, 30$. Clearly, test sTREM-1 is highly unlikely to meet the joint criteria set for the three groups, while test IL-6 does considerably better but the NPI lower and upper probabilities remain quite low. There is much more imprecision in the values for IL-6 than for sTREM-1, this is simply due to the fact that the NPI upper probabilities for the latter are quite close to 0.

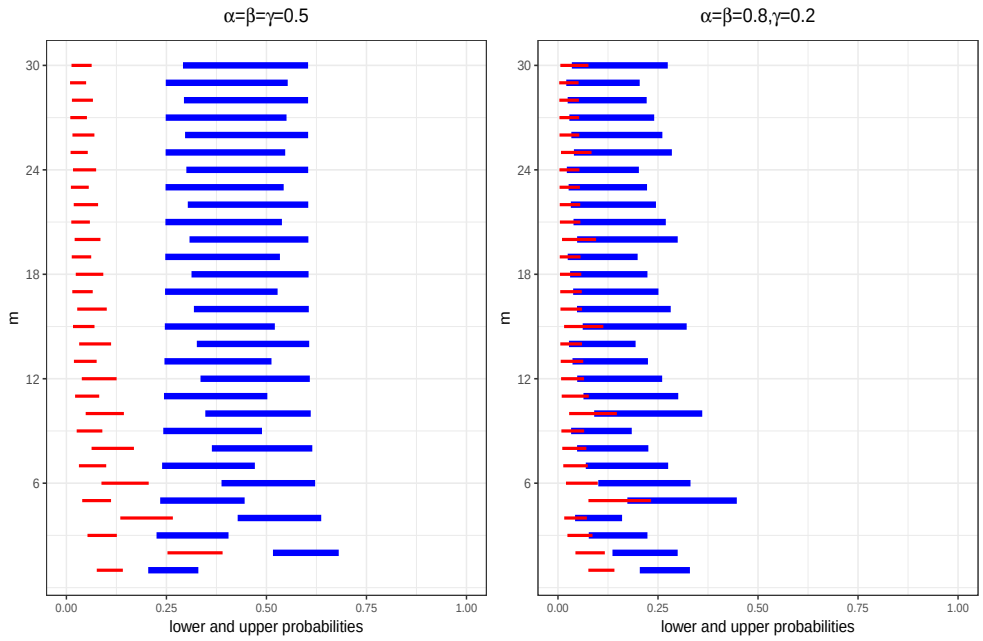

Figure 4: Comparison of IL-6 (blue, thick) and sTREM-1 (red, thin).

As the performance of the test sTREM-1 is poor for group Z, we also illustrate our method for the scenario with $\alpha = \beta = 0.8, \gamma = 0.2$. The optimal test thresholds based on the 3-NPI-L method (Alabdulhadi, 2018; Coolen-Maturi et al., 2020), with these required proportions of successful future diagnoses, lead to the following empirical results. For test IL6 the empirical results are identical to those reported above for the case with all three minimum required proportions equal to 0.5. For test sTREM-1, the numbers of correctly classified individuals from groups X, Y and Z are 16 out of 21, 5 out of 9 and 6 out of 22 respectively, for $m = 1, 2, 6, 11$, which are the same numbers as above, but for the other considered values of m these numbers are 16 out of 21, 7 out of 9 and 2 out of 22 respectively. Note that this latter case has a substantially lower number of data from the Z group correctly diagnosed, which is in line with the low criterion now set for that group. However, it should be noted that, when considering the empirical data, both tests fail to meet the criteria set for all three groups together.

The predictive comparison using our new method, as presented in this section, is based on the lower and upper probabilities shown in the second plot in Figure 4. For most considered values of m, there is a weak indication that test IL-6 is better than test sTREM-1, according to this predictive requirement. However, for $m = 1, 2$, there is a strong indication that test IL-6 is better than test sTREM-1. It should be noted that, in both these plots, we see again the effect of the discrete nature of the numbers of future individuals that must be correctly diagnosed in the three groups, which e.g. implies that the criteria with $\alpha = \beta = 0.8, \gamma = 0.2$ are easier achieved for $m = 5$ than for $m = 4$, as in both cases 4 out of m future individuals

from groups X and Y must be correctly classified, together with 1 out of m for group Z.

5. Concluding remarks

This paper presents comparison of two diagnostic tests by explicitly focussing on their predictive performance when applied to future individuals. The theory was developed for possibly different numbers of future individuals from the two or three groups. While we illustrated the method in the examples with these numbers assumed to be equal, the flexibility of the method may be an advantage for practical application. Specific choice of these numbers is left as a topic for future research, two ideas worth considering are as follows. One may have an idea about the likely numbers of people from each group to present themselves for diagnoses over a given future time period, in which these numbers can be used to reflect the expected practical circumstances. An alternative choice could be to set $m_i = n_i$ for each $i = x, y, z$. This could provide insight into reproducibility of the overall diagnostic performance of the test as indicated empirically by using the available data. This is in line with the recently presented use of NPI for reproducibility of statistical tests (Coolen and Alqifari, 2018; Coolen and Bin Himd, 2014) and this is an interesting topic for future research.

The method presented in this paper requires choice of the minimum proportions of correct diagnoses for future individuals from each group, α, β and γ. This provides a simple way to take the importance of correct diagnosis for each group into account. Of course, one could argue that these values should be close to 1, but this may lead to extremely small values for the NPI lower and upper probabilities of the event that the criteria will be met for each of the two or three groups. If these proportions are set very low, on the other hand, these NPI lower and upper probabilities will be very large, which is also unlikely to provide useful insights into the performance of the tests. Further research is required into sensible choice of these proportions. One could, for example, set them close to the empirical proportions, in order to get NPI lower and upper probabilities which are not too close to 0 or 1. The main idea of using these proportions has been to provide a simple predictive criterion for the comparison of the two tests. It is also possible to use different predictive criteria, for example the use of the (possibly weighted) sum of correctly classified individuals for all groups has also been studied in the recent PhD thesis of the first-named author (Alabdulhadi, 2018), and it is an interesting topic for future research to consider more predictive criteria for successful performance of diagnostic tests and to use these to compare such tests.

Acknowledgement

We are grateful to Prof. Sarafidis for providing the data set used in Example 4.1. The authors are grateful to the anonymous reviewer whose supportive comments led to improved presentation of the paper.

References

Alabdulhadi, M.H. (2018). *Nonparametric Predictive Inference for Diagnostic Test Thresholds*. PhD thesis, Durham University (available from *www.npi-statistics.com*.)

Alqifari, H.N. (2017). *Nonparametric Predictive Inference for Future Order Statistics*. PhD thesis, Durham University (available from *www.npi-statistics.com.*)

Augustin, T. and Coolen, F.P.A. (2004). Nonparametric predictive inference and interval probability. *Journal of Statistical Planning and Inference*, **124**, 251-272.

Augustin, T., Coolen, F.P.A., de Cooman, G. and Troffaes, M.C.M. (2014). *Introduction to Imprecise Probabilities*. Wiley, Chichester.

Coolen, F.P.A. (2011). Nonparametric predictive inference, In: *International Encyclopedia of Statistical Science*, Lovric, M. (ed.). Springer, Berlin, pp. 968-970.

Coolen, F.P.A. and Alqifari, H.N. (2018). Nonparametric predictive inference for reproducibility of two basic tests based on order statistics. *Revstat - Statistical Journal*, **16**, 167-185.

Coolen, F.P.A. and Bin Himd, S. (2014). Nonparametric predictive inference for reproducibility of basic nonparametric tests. *Journal of Statistical Theory and Practice*, **8**, 591-618.

Coolen, F.P.A., Coolen-Maturi, T. and Alqifari, H.N. (2018). Nonparametric predictive inference for future order statistics. *Communications in Statistics - Theory and Methods*, **47**, 2527-2548.

Coolen-Maturi, T., 2017a. Three-group ROC predictive analysis for ordinal outcomes. *Communications in Statistics: Theory and Methods*, **46**, 9476-9493.

Coolen-Maturi, T., 2017b. Predictive inference for best linear combination of biomarkers subject to limits of detection. *Statistics in Medicine*, **36**, 2844-2874.

Coolen-Maturi, T., Coolen, F.P.A. and Alabdulhadi, M.H. (2020). Nonparametric predictive inference for diagnostic test thresholds. Communications in Statistics: Theory and Methods, 49 (3): 697–725.

Coolen-Maturi, T., Coolen-Schrijner, P., Coolen, F. P., 2012a. Nonparametric predictive inference for diagnostic accuracy. Journal of Statistical Planning and Inference 142 (5), 1141–1150.

Coolen-Maturi, T., Coolen-Schrijner, P., Coolen, F. P., 2012b. Nonparametric predictive inference for binary diagnostic tests. Journal of Statistical Theory and Practice 6 (4), 665–680.

Coolen-Maturi, T., Elkhafifi, F. F., Coolen, F. P., 2014. Three-group roc analysis: A nonparametric predictive approach. Computational Statistics & Data Analysis 78, 69–81.

Cox, L.H., Johnson, M.M. and Kafadar, K. (1982). Exposition of statistical graphics technology. *ASA Proceedings of the Statistical Computation Section*, 55-56.

Elkhafifi, F.F. and Coolen, F.P.A. (2012). Nonparametric predictive inference for accuracy of ordinal diagnostic tests. *Journal of Statistical Theory and Practice*, **6**, 681-697.

Hill, B. M., 1968. Posterior distribution of percentiles: Bayes' theorem for sampling from a population. Journal of the American Statistical Association 63 (322), 677–691.

Hill, B. M., 1988. De finetti's theorem, induction, and A(n) or bayesian nonparametric predictive inference (with discussion). Bayesian statistics 3, 211–241.

Pepe, M. S., 2003. The Statistical Evaluation of Medical Tests for Classification and Prediction. Oxford University Press, Oxford.

Sarafidis, K., Soubasi-Griva, V., Piretzi, K., Thomaidou, A., Agakidou, E., Taparkou, A., Diamanti, E. and Drossou-Agakidou, V. (2010). Diagnostic utility of elevated serum soluble triggering receptor expressed on myeloid cells (sTREM)-1 in infected neonates. *Intensive Care Medicine*, **36**, 864-868.

Weichselberger, K., 2000. The theory of interval-probability as a unifying concept for uncertainty. International Journal of Approximate Reasoning 24 (2), 149–170.

Zhou, X.H., Obuchowski, N.A. and McClish, D.K. (2002). *Statistical Methods in Diagnostic Medicine*. Wiley, New York.